2-3

УДК 004.421

Т.М. Заболотня, А.Ю. Михайлюк, О.С. Михайлюк
Національний технічний університет України «Київський політехнічний інститут»,
м. Київ, Україна
tatiana104@yandex.ru

Інверсний контекстно-асоціативний метод автоматизованої орфокорекції

Теоретично обґрунтовано та запропоновано інверсний контекстно-асоціативний метод автоматизованого виправлення орфографічних помилок, який забезпечує підвищення швидкості та точності роботи відповідного програмного забезпечення. Дано визначення показника результативності функціонування орфокоректора – точності його роботи. Показана ефективність застосування запропонованого методу для виправлення орфографічних помилок у масиві гетерогенних словосполучень за критеріями швидкості та точності корекції.

Вступ

На сучасному етапі розвитку суспільства, в умовах зростання потреби у підвищенні рівня інтелектуалізації інформаційних технологій великого значення набула проблема забезпечення ефективної комп'ютеризованої обробки природномовних текстів.

Важливою задачею систем автоматичної обробки текстів (АОТ) різного призначення є перевірка орфографії в текстових даних. На даний момент актуальним є дослідження програмного забезпечення орфокорекції з точки зору виправлення помилок, адже задача виявлення помилок досліджувалась триваліший час і розв'язана у більшій мірі [1].

Сучасні орфокоректори характеризуються невисоким рівнем автоматизації, орієнтацією на виправлення однократних помилок, відсутністю забезпечення семантичної узгодженості варіантів виправлення та контексту спотвореного слова [1], [2]. Ці недоліки зумовлюють низьку ефективність роботи коректорів, основним критерієм оцінки якої на сьогоднішній день виступають точність та швидкість обробки текстових даних.

Усунення визначених вище вад потребує модифікації існуючих алгоритмів виправлення помилок у напрямку залучення контекстної інформації та лексико-семантичних словникових ресурсів для її обробки. Але на даний момент в програмних засобах орфокорекції методика врахування контексту спотвореного слова за допомогою семантичного інструментарію є недостатньо розробленою, а отже, дослідження даного питання видається актуальним.

1. Постановка задачі

На сьогоднішній день розробка автокоректорів є окремим широко досліджуваним напрямом у галузі створення систем АОТ. Але, незважаючи на це, способи побудови ефективних програмних засобів корекції спотворених слів здебільшого визначаються загальними характеристиками сучасного етапу розвитку систем АОТ та комп'ютерної техніки взагалі.

1. З огляду на постійне зростання обсягів пам’яті та швидкодії сучасних комп’ютерів основними вимогами до систем АОТ сьогодні є точність та швидкість їх роботи. Це сприяє все більшому розповсюдженню використання словників при аналізі даних, яке раніше було неможливим через труднощі, зумовлені підвищеними вимогами до обсягу пам’яті в запам’ятовуючих пристроях. Словниковий підхід забезпечує отримання високоточних результатів обробки текстів [2], [3]. Точність орфокорекції у наукових джерелах визначається як відсоток спотворених слів, для котрих програмою підібрано вірний варіант виправлення [4], або як імовірність входження правильного варіанта до набору гіпотез [5].

2. На відміну від програмного забезпечення кін. ХХ ст., коли основу машинної обробки тексту складав морфологічний аналіз, теперішні системи широко використовують ресурси та методи синтаксичного та семантичного аналізу (для автоматичного анотування, реферування, машинного перекладу тощо).

3. Для забезпечення можливості врахування значення слів системами АОТ у більшості випадків розробники віддають перевагу аналізу контексту, а не семантики граматичної структури окремих слів [6].

4. Загальноприйнятою схемою аналізу тексту залишається так звана послідовна схема (морфологічний, синтаксичний, семантичний рівні аналізу), коли результати кожного попереднього етапу є вихідною інформацією для наступних [6].

Можна сказати, що всі наведені характеристики систем АОТ властиві сучасним автокоректорам у тій чи іншій мірі [5], [7], [8]. Особливу увагу слід приділити схемі аналізу текстових даних, якої дотримуються у своїй роботі системи АОТ. Справа у тому, що етапи аналізу природномовного тексту не є функціонально ізольованими, хоч і виконуються зазвичай послідовно. Згідно з цим, морфологічний аналіз може не лише надавати вихідні дані для синтаксичного та семантичного аналізу, але і використовувати результати останніх [9]. Така тенденція до порушення класичного порядку аналізу текстів сьогодні вже спостерігається у системах АОТ [1], [9]. Але досі обробка контексту спотвореного слова семантичним інструментарієм має місце тільки на кінцевому етапі процесу комп’ютеризованої орфокорекції.

Таким чином, виходячи із наведених вище аргументів, **метою даної статті** стало підвищення швидкості та точності автоматизованого виправлення орфографічних помилок відповідними програмними засобами за рахунок створення контекстно-асоціативного методу корекції спотворених слів.

У відповідності до поставленої мети **задачами дослідження** є:

– вивчення способів реалізації етапів процедури орфокорекції на предмет виявлення можливостей щодо підвищення її ефективності;

– розробка контекстно-асоціативного методу орфокорекції для підвищення ефективності роботи відповідного програмного забезпечення за показниками точності та швидкості виправлення помилок;

– аналіз результатів експериментального дослідження ефективності запропонованого контекстно-асоціативного методу орфокорекції.

2. Схема автоматизованого виправлення орфографічних помилок

Загальноприйнята схема автоматизованої корекції спотвореного слова [1], [10] передбачає реалізацію:

– етапу висунення гіпотез (вірогідних варіантів виправлення помилки);

– етапу перевірки гіпотез та ухвалення однієї (декількох) з них як виправлення, що пропонується програмою до внесення.

На першому етапі послідовно виконуються підбір первинної множини варіантів виправлення із словника та попередня фільтрація її вмісту. Для цього використовуються *найпростіші* та *найшвидші* методи пошуку варіантів корекції слова (наприклад, підбір гіпотез за критерієм альфакоду, довжини слова, збігу першої літери слова [10] тощо).

На другому етапі виконується перевірка гіпотез на подібність до спотвореного слова за певними критеріями. Тут задіяні більш *складні*, але, водночас, і більш *точні* методи аналізу набору гіпотез (наприклад, відстань редагування Левенштейна).

Таким чином, умовне віднесення методів визначення варіантів виправлення орфографічних помилок до певного етапу процесу орфокорекції здійснюється на основі їх характеристик (швидкості, точності тощо).

З іншого боку, методи висунення та перевірки гіпотез виправлення за своєю суттю передбачають фільтрацію заданої множини слів, адже в результаті застосування кожного з них відбувається звуження поточної множини варіантів корекції спотвореного слова. З огляду на це введемо функцію фільтрації множини слів за певною ознакою *filter* та визначимо її властивості.

Визначення 1. Функція $filter: W_x \to W_y$ називається фільтром множини W_x, якщо за її допомогою з елементів W_x проводиться формування множини слів W_y, які відповідають певному критерію схожості зі спотвореним словом ($W_y \subseteq W_x$).

$$filter: W_x \to W_y,\ W_y \subseteq W_x, \tag{1}$$

де W_x, W_y – множини природномовних слів.

Виходячи з фізичного змісту функції *filter*, її властивостями є:

1) адитивність: $filter(W_A \cup W_B) = filter(W_A) \cup filter(W_B)$; (2)

2) комутативність композиції фільтрів: при застосуванні композиції функцій $F = filter_n \circ filter_{n-1} \circ \ldots \circ filter_2 \circ filter_1 : W_x \to W_y,\ W_y \subseteq W_x$ до множини слів W_x від перестановки складових $filter_i$ місцями результат W_y не змінюється, адже у будь-якому випадку всі слова, які не відповідають хоча б одному критерію відбору, будуть вилучені з множини W_x;

3) якщо $W_A \subseteq W_B$, то час, необхідний для виконання фільтрації даних множин, характеризується нерівністю

$$t_{filter(W_A)} \le t_{filter(W_B)}, \tag{3}$$

де $t_{filter(W_A)}, t_{filter(W_B)}$ – час фільтрації множин W_A, W_B за допомогою функції *filter*. Оскільки всі функції, які реалізують методи відбору та перевірки гіпотез виправлення, є фільтрами, їм притаманні властивості визначеної вище функції *filter*.

Виходячи з вищенаведеного, пропонується внести уточнення в подання схеми орфокорекції: будемо вважати процес визначення варіантів виправлення таким, що складається із застосування композиції функцій фільтрації до множини слів W_{dict}, яка міститься у словнику. Позначимо послідовність фільтрів як

$$FILTERS = f_m \circ f_{m-1} \circ \ldots \circ f_i \circ \ldots \circ f_2 \circ f_1 : W_{dict} \to W_{retr},\quad m > 1, \tag{4}$$

де $f_i : W_{i-1} \to W_i$ (i = 2, …, m) – фільтр множини слів, отриманої у результаті виконання f_{i-1} (для f_1 – множини W_{dict}); W_{retr} – множина слів, визначених коректором як можливі варіанти виправлення за ознаками їх близькості до спотвореного слова.

Перш ніж приступити до пошуку місця семантичної складової у схемі орфокорекції, реалізація якої забезпечує найбільш ефективну роботу програмного коректора за критеріями точності та швидкості, зупинимось на визначенні точності результату виправлення помилок.

Сучасні словникові ресурси характеризуються великим обсягом даних [6], що ускладнює однозначний підбір з них правильного варіанта написання спотвореного слова. Тому виправданим способом оцінювання точності орфокорекції є її обчислення як частки спотворених слів, до множини варіантів виправлення яких входить вірна словоформа [5]. Але для забезпечення високого рівня автоматизації обробки текстових даних потрібно не тільки збільшення імовірності входження правильного варіанта виправлення до набору гіпотез, але й зменшення кількості знайдених нерелевантних гіпотез. Через це необхідним є знаходження способу оцінювання ефективності орфокорекції за критерієм точності, який забезпечив би врахування частки нерелевантних слів у множині варіантів виправлення.

У даній статті йдеться про визначення гіпотез шляхом їх *пошуку* в словнику (а не за допомогою безсловникової генерації), тому при визначенні точності орфокорекції пропонується провести певні паралелі із оцінками результатів роботи програм у теорії інформаційного пошуку [6].

Визначення 2. Під *точністю* машинної орфографічної корекції спотвореного слова будемо розуміти відношення числа запропонованих орфокоректором вірних варіантів написання слова (це одиниця або нуль) до загальної кількості підібраних слів.

$$PRECISION = \frac{|W_{corr} \cap W_{retr}|}{|W_{retr}|}, \qquad (5)$$

де W_{corr} – множина вірних варіантів корекції спотвореного слова у словнику.

Відповідно до формули (5), для того, щоб досягти високого показника точності роботи орфокоректора, необхідно, по-перше, забезпечити постійне входження вірного слова до сформованого масиву варіантів виправлення ($|W_{corr} \cap W_{retr}| = 1$), а по-друге, – зменшити загальну кількість слів, які пропонуються програмою як найбільш вірогідні кандидати виправлення помилки (W_{retr}).

3. Місце семантичної складової у схемі виправлення орфографічних помилок

Згідно з класичною послідовністю обробки текстових даних (морфологічний, синтаксичний та семантичний аналіз), семантичні фільтри набору гіпотез мають стояти наприкінці композиції *FILTERS* (4). Відповідно ж до сучасних тенденцій щодо зміни загальноприйнятого порядку етапів обробки текстів, можливим є підвищення ефективності програмного орфокоректора у випадку перенесення перевірки гіпотез за семантичними критеріями ближче до початку *FILTERS*. Проведемо дослідження впливу зміни місця семантичної складової у схемі орфокорекції на показники точності та швидкодії програмного коректора.

Формування множини гіпотез виправлення за семантичним критерієм із заданого набору слів здійснюватимемо за допомогою функції f_{cont}. Визначимо дану функцію як фільтр, який застосовується для відбору із вихідного набору слів тих словоформ, що узгоджені з контекстним оточенням спотвореного слова.

Розглянемо 3 варіанти розміщення f_{cont} у схемі визначення варіантів виправлення:

1) контекстно-асоціативну фільтрацію гіпотез як останній етап процесу підбору варіантів виправлення:

$$f_{cont} \circ f_m \circ f_{m-1} \circ ... \circ f_2 \circ f_1 : W_{dict} \to W_{retr},\ m >= 1, \qquad (6)$$

2) проведення відбору гіпотез за ознакою їх відповідності змісту заданого контексту в середині процесу орфокорекції:

$$f_m \circ f_{m-1} \circ ... \circ f_{cont} \circ f_{i-1} \circ ... \circ f_2 \circ f_1 : W_{dict} \to W_{retr},\ m >= 1, \qquad (7)$$

3) висунення гіпотез за критерієм їх семантичної близькості до контексту спотвореного слова (перенесення f_{cont} на початок схеми корекції):

$$f_m \circ f_{m-1} \circ ... \circ f_2 \circ f_1 \circ f_{cont} : W_{dict} \to W_{retr},\ m >= 1. \qquad (8)$$

Твердження 1. Зміна розташування фільтра f_{cont} у схемі орфокорекції не впливає на точність роботи коректора (*PRECISION*).

Доведення. Дане твердження є коректним, виходячи з властивості (2) функцій класу *filter*: перестановка фільтрів місцями не приводить до зміни результуючої множини слів, а отже, і рівень точності не змінюється, що і потрібно було довести.

Тепер визначимо, при якому з трьох наведених варіантів розміщення f_{cont} у схемі визначення варіантів виправлення (6) – (8) і за яких умов буде досягатися найвища швидкодія орфокоректора.

Розглянемо спочатку застосування семантичного фільтра в кінці процесу підбору варіантів виправлення та у його середині (6), (7) .

Нехай W_{i-1} – результат фільтрації множини W_{dict} із використанням композиції функцій $f_{i-1} \circ ... \circ f_2 \circ f_1 : W_{dict} \to W_{i-1}$ (для $i = 1$ роль W_{i-1} виконує безпосередньо W_{dict}). Для обох випадків, що аналізуються, вміст W_{i-1} є однаковим, адже вихідна множина гіпотез і набір функцій, які до неї застосовуються, не відрізняються.

Таким чином, для того, щоб швидкодія коректора, який реалізує послідовність фільтрів (7), була не нижчою за швидкість його роботи за схемою (6), має виконуватися нерівність:

$$t_{f_{cont}(W_{i-1})} + \sum_{k=i}^{m} t_{f_k(Wcont_{k-1})} <= \sum_{k=i}^{m} t_{f_k(W_{k-1})} + t_{f_{cont}(W_m)}, \qquad (9)$$

де $Wcont_{i-1}$ – результат фільтрації слів з W_{i-1} за ознакою близькості за змістом до контекстного оточення спотвореного слова. Проаналізуємо, у якому випадку дана нерівність буде справедливою.

f_{cont} за визначенням є фільтром, тому справедливим є твердження $f_{cont} : W_{i-1} \to Wcont_{i-1}$, $Wcont_{i-1} \subseteq W_{i-1}$. Отже, маємо:

$$Wcont_{i-1} \cup \Delta W_{cont_out} = W_{i-1}, \qquad (10)$$

де ΔW_{cont_out} – частина множини W_{i-1}, яка була виключена із подальшої обробки через невідповідність семантичному критерію фільтрації слів.

Перевірка множин W_{i-1} та $Wcont_{i-1}$ за допомогою функцій, які входять до складу композицій (6) та (7) відповідно, проводиться, починаючи з фільтра f_i:

– $W_{i-1} \xrightarrow{f_i} W_i,\ W_i \subseteq W_{i-1}$;

– $Wcont_{i-1} \xrightarrow{f_i} Wcont_i,\ Wcont_i \subseteq Wcont_{i-1}$.

Виходячи з того, що $Wcont_{i-1} \subseteq W_{i-1}$, можна стверджувати: згідно з (3), нерівність $t_{f_i(Wcont_{i-1})} <= t_{f_i(W_{i-1})}$ є вірною.

Відповідно до (2) та (10):

$$f_i(W_{i-1}) = f_i(Wcont_{i-1}) \cup f_i(\Delta W_{cont_out}) \Rightarrow,$$
$$W_i = Wcont_i \cup f_i(\Delta W_{cont_out}) \Rightarrow Wcont_i \subseteq W_i .$$

Застосування фільтра f_{i+1} характеризується аналогічним чином:

$$W_i \xrightarrow{f_{i+1}} W_{i+1}, \; W_{i+1} \subseteq W_i \text{ та } Wcont_i \xrightarrow{f_{i+1}} Wcont_{i+1}, \; Wcont_{i+1} \subseteq Wcont_i .$$

Звідки

$$f_{i+1}(W_i) = f_{i+1}(Wcont_i) \cup f_{i+1} \circ f_i(\Delta W_{cont_out}) \Rightarrow W_{i+1} = Wcont_{i+1} \cup f_{i+1} \circ f_i(\Delta W_{cont_out}) \Rightarrow$$
$$Wcont_{i+1} \subseteq W_{i+1} \text{ і т.д.}$$

У підсумку на основі властивості 3) (3) функції *filter* отримуємо

$$\sum_{k=i}^{m} t_{f_k(Wcont_{k-1})} <= \sum_{k=i}^{m} t_{f_k(W_{k-1})}.$$

Тому для того, щоб перенесення семантичного фільтра ближче до початку послідовності *FILTERS* сприяло прискоренню роботи орфокоректора, у випадку, який розглядається, необхідно, щоб виконувалась умова $t_{f_{cont}(W_{i-1})} <= t_{f_{cont}(W_m)}$ (9) .

Якщо аналогічним чином провести аналіз умов підвищення швидкодії програмних засобів орфокорекції для пар композицій функцій (6) – (8), (7) – (8), можна дійти висновку, що при перенесенні семантичного фільтра ближче до початку послідовності *FILTERS* коректор працюватиме швидше за умови виконання нерівностей $t_{f_{cont}(W_{dict})} <= t_{f_{cont}(W_m)}$ та $t_{f_{cont}(W_{dict})} <= t_{f_{cont}(W_{i-1})}$ відповідно.

Таким чином, для того, щоб обрати варіант розміщення f_{cont} у схемі визначення варіантів виправлення, який забезпечує найвищу швидкодію орфокоректора, необхідно порівняти час виконання перевірки гіпотез на відповідність контексту за допомогою функції f_{cont} у кожному із згаданих випадків.

W_m є результатом послідовної фільтрації W_{i-1}, яка в свою чергу отримується шляхом відбору слів з множини W_{dict} ($W_m \subseteq W_{i-1} \subseteq W_{dict}$). Тому, згідно з визначенням функції *filter* (3), виконання вищезгаданих нерівностей є неможливим, якщо не існує жодної відмінності у реалізації f_{cont} при зміні її місця у схемі орфокорекції.

Отже, для визначення існування можливості підвищення швидкодії орфокоректора при переміщенні f_{cont} у послідовності фільтрів вмісту словника необхідно дослідити специфіку реалізації даної семантичної функції на різних етапах схеми орфокорекції. Оскільки послідовності фільтрів (6) та (7) є подібними (в першому випадку функції f_{cont} передує композиція з *m* фільтрів, а в другому – з *i* фільтрів), далі аналізувати будемо тільки два варіанти розміщення функції f_{cont}: прямий порядок фільтрації гіпотез (6) та інверсний (8).

Будемо вважати, що:

- час, потрібний для вибору із словника одного слова за будь-якою ознакою, є однаковим для усіх критеріїв відбору в межах задачі орфокорекції;
- час, потрібний для перевірки слова на відповідність будь-якій ознаці, є однаковим для усіх простих формальних фільтрів;
- будь-яка вибірка слів, підібраних за певним критерієм, зберігає репрезентативність відносно інших критеріїв фільтрації.

Крім цього, *мірою семантичної близькості двох слів* вважатимемо довжину шляху між відповідними вершинами графа словника, а кількісне оцінювання *семантичної близькості альтернативи виправлення до контекстного оточення спотвореного слова* виконуватимемо на основі визначення величини, оберненої мінімальній з довжин найкоротших шляхів від заданого слова до контексту за структурою словника [11]. Для забезпечення орфокоректора даними про семантично-асоціативні зв’язки між словами природної мови будемо використовувати онтологічний словниковий ресурс у формі орієнтованого графу, вершинами якого є лексеми природної мови, поєднані лексико-семантичними відношеннями [9], [11].

З огляду на те, що із збільшенням дистанції між вершинами графа словника сила семантичного зв’язку між ними швидко зменшується [9], слова, відстань від яких до контексту перевищує певний поріг *maxdist*, будемо вважати нескінченно віддаленими від нього і не включатимемо їх до множини гіпотез. Це дасть змогу отримувати набір варіантів виправлення, котрі мають задану міру семантичної близькості до контексту, аналізуючи при цьому обмежену частину словникового ресурсу (оточення контекстних слів у радіусі *maxdist*). При цьому процедуру визначення слів, близьких за змістом до контексту, можна організувати таким чином, що для її успішного завершення достатньо обробити вершини графа словника у радіусі $R_{\min}$ від контексту, де:

$$R_{\min} = \min(R)\text{, де } R = \{r \in \mathrm{N} \mid (r <= \mathrm{maxdist}) \wedge (((\bigcup_i \sigma x_i^{\,r} \cup \sigma x_i^{\,-r}) \cap W_x) \neq \varnothing)\} \qquad (11а)$$

$$\text{або } R = \{r \in \mathrm{N} \mid (r <= \mathrm{maxdist}) \wedge (F_A(\bigcup_i \sigma x_i^{\,r} \cup \sigma x_i^{\,-r}) \neq \varnothing)\}, \qquad (11б)$$

$W_x \subset W_{dict}$ – результат попереднього відбору гіпотез, які потрібно перевірити на семантичну відповідність контексту; F_A – абстрактна функція перевірки елементів заданої множини слів на відповідність іншим критеріям схожості із спотвореним словом; x_i – слово контекстного оточення; $\text{у}x_i^{\,r}, \text{у}x_i^{\,-r}$ – відображення *r*-го ступеня вершини x_i графа словника (пряме та зворотне).

Твердження 2. Застосування семантичної функції f_{cont} під час висунення гіпотез виправлення забезпечує більш швидке отримання результату роботи орфокоректора, ніж її використання для остаточної перевірки множини гіпотез.

Доведення. Будемо аналізувати вершини графа словника, які лежать у радіусі $R_{\min}$= *maxdist* від слів контексту (щоб розглянути випадок, коли необхідною є обробка максимально припустимої кількості слів, семантично пов’язаних з контекстом спотвореного слова). Введемо позначення:

– $|context|$ – кількість слів контексту, які містяться у словнику;

– *y* – максимальна кількість лексико-семантичних зв’язків (дуг графа), які має одне слово $w \in W_{dict}$ з іншими лексемами словника; *y* має порядок, близький до порядку величини $|context|$.

а) *Інверсна* послідовність фільтрів (8). Окіл контексту *context* із радіусом *maxdist* становить $A = |\,context\,| * \sum_{j=1}^{\mathrm{maxdist}} y^j$ слів [11]. Під час виконання $f_{cont}(W_{dict})$ із словника відбирається *A* слів і передається для подальшої фільтрації без проведення додаткових перевірок.

б) *Пряма* послідовність фільтрів (6). Для перевірки семантичної узгодженості гіпотез та контексту спотвореного слова функції f_{cont} передається $|W_m|$, $W_m \subseteq W_{dict}$ слів. Разом з цим для роботи фільтра $f_{cont}(W_m)$ у будь-якому випадку необхідно знати окіл контексту (у даному доведенні – це *A* слів), а отже, додатково потрібно отримати *A* слів із словника. Виконання $|W_m| * A$ додаткових перевірок закінчує роботу даної послідовності фільтрів.

Таким чином, інверсний порядок виконання фільтрації спричиняє виконання меншої кількості дій і, значить, забезпечує більш швидку роботу функції f_{cont} (тобто $t_{f_{cont}(W_{dict})} <= t_{f_{cont}(W_m)}$). Вище показано, що за умови справедливості даної нерівності можна стверджувати, що час виконання композиції фільтрів (8) є меншим, ніж час виконання композиції (6), що і потрібно було довести. (Для випадків, коли $R_{\min} < maxdist$, справедливість зазначених нерівностей зберігається.)

4. Інверсний контекстно-асоціативний метод автоматизованого виправлення орфографічних помилок

З огляду на викладені вище результати дослідження загальної схеми орфокорекції та особливостей її реалізації можна визначити *інверсний контекстно-асоціативний метод автоматизованої орфокорекції.* В основу методу пропонується покласти встановлення зворотного порядку фільтрації вмісту словника, оскільки доведено факт підвищення швидкості пошуку варіантів виправлення при збереженні точності роботи орфокоректора у разі перенесення семантичної складової схеми корекції на її початковий етап.

Згідно з даним методом процедура контекстно-асоціативної автоматизованої орфокорекції являє собою послідовне виконання таких дій:

1) встановлення радіуса пошуку r рівним мінімально припустимому значенню;
2) висунення гіпотез виправлення за ознакою семантичної близькості до контекстного оточення спотвореного слова;
3) перевірка гіпотез виправлення на подібність до спотвореного слова за формальними ознаками;
4) збільшення радіуса пошуку гіпотез виправлення та перехід до п. 2 даного методу у випадку, якщо, по-перше, $r < maxdist$, а по-друге, якщо на заданій відстані r від вершин графа словника, котрі відповідають словам контексту, не знайдено жодного слова, яке задовольнило б усім критеріям схожості зі спотвореним словом; в іншому разі – закінчення пошуку варіантів виправлення.

Важливою особливістю методу є його ітераційний характер. Він дозволяє зменшити кількість дій щодо обробки слів під час орфокорекції і тим самим підвищити швидкість її виконання.

5. Дослідження ефективності інверсного контекстно-асоціативного методу орфокорекції

Для експериментальної апробації запропонованого методу орфокорекції, а також для перевірки справедливості теоретичних тверджень, які доводяться у статті, використано масив словосполучень, які характеризуються різною потужністю множин слів, які складають контекст спотвореного слова; різною кількістю помилок, припущених у слові (1 та 2 помилки); різною силою семантичного зв’язку контексту із спотвореним словом.

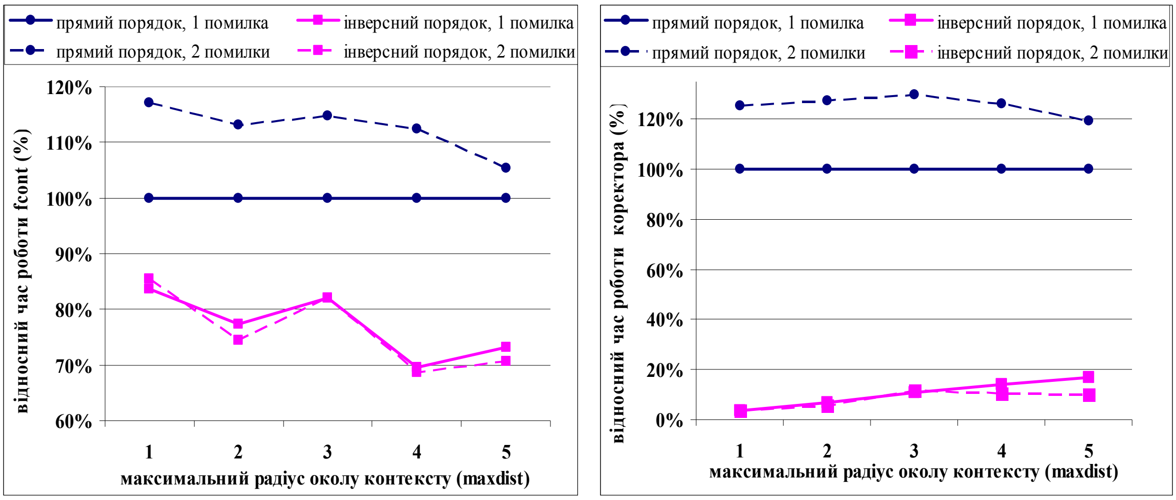
Для підтвердження досягнення найкращих показників роботи коректора за умови застосування фільтрів до вмісту лексико-семантичного словника в *інверсному* порядку проаналізовано результати функціонування відповідного програмного забезпечення у випадках, коли алгоритмом його роботи передбачено:

1) використання семантичної функції f_{cont} *на початку* та *наприкінці* послідовності фільтрів (*прямий* та *інверсний* порядок фільтрації);

2) проведення спроб виправлення *одно*- та *двократних* помилок;

3) встановлення радіуса околу контекстного оточення спотвореного слова за структурою графа словника рівним від 1 до 5 переходів (*maxdist* = 1...5).

Крім того, показники роботи орфокоректора порівнюються із аналогічними показниками відповідного модуля, вбудованого до пакета MS Word, функціональність якого сьогодні найчастіше використовується для обробки текстів.



а б

Рисунок 1 – Графік залежності відносного часу роботи семантичної функції (а) та орфокоректора взагалі (б) від особливостей реалізації алгоритму орфокорекції

Різноманітність даних, які надходять на вхід орфокоректора, робить неможливим отримання узагальненої оцінки часу корекції в секундах (чи інших одиницях виміру часу). Тому результати вимірювань часових значень наводяться відносно відповідних показників роботи коректора, коли він виконує виправлення *однократних* помилок із застосуванням фільтрів вмісту словника в *прямому* порядку.

Як бачимо на рис. 1а, проведення фільтрації вмісту словникового ресурсу в інверсному порядку приводить до покращення часових характеристик роботи функції f_{cont} (порівняно із застосуванням f_{cont} в кінці фільтрації) незалежно від радіуса пошуку варіантів виправлення у графі словника.

При застосуванні фільтрів у прямому порядку час виконання f_{cont} при виправленні двократних помилок є більшим, ніж час її роботи при корекції слів з одинарними помилками; при застосуванні фільтрів в інверсному порядку дані часові показники практично збігаються.

На рис. 1 не подано характеристики роботи MS Word, оскільки час його роботи можна порівнювати з часом роботи розробленого орфокоректора тільки за умови аналізу околу контексту з радіусом *R = maxdist*. У разі *maxdis t*<= 4 швидкість роботи MS Word значно (до 3 разів) поступається швидкодії нового орфокоректора.

Отже, умова $t_{f_{cont}(W_{dict})} <= t_{f_{cont}(W_m)}$ (п. 3) виконана, і загальний час проведення орфокорекції повинен бути меншим при застосуванні фільтрів до вмісту словникового ресурсу в інверсному порядку, що і підтверджується рис. 1б.

Практичні дослідження підтвердили справедливість теоретично прийнятої комутативності композиції функцій *filter*. Разом з тим, очевидною є перевага розроблених програмних засобів орфокорекції за критерієм точності над модулем

виправлення помилок MS Word (рис. 2). Найімовірнішою причиною таких результатів є припущення про те, що останній не реалізує перевірки варіантів виправлення на семантичну узгодженість з контекстом спотвореного слова [2].

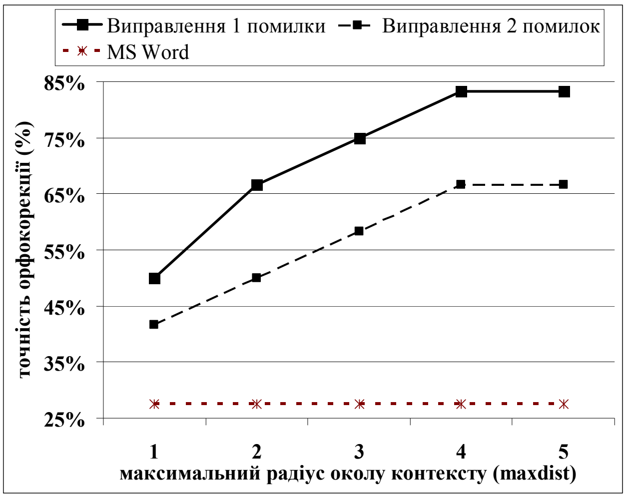


Рисунок 2 – Графік залежності точності роботи програмних засобів від реалізації алгоритму орфокорекції

Факт, що розроблене ПЗ, яке реалізує запропонований метод автоматизованої орфокорекції, не характеризується 100 % точністю роботи, пояснюється тим, що, з одного боку, при виправленні однократних помилок слова, які мають двократні помилки, залишаються невиправленими, а з іншого боку, спроба виправлення двократних помилок спричиняє додавання до результатів зайвих слів у випадку корекції слів, які мають однократну помилку.

Можна відмітити ще одну важливу закономірність: точність виправлення помилок перестає зростати при *maxdist* > 4. Отже, з точки зору досягнення найвищої точності роботи орфокоректора збільшувати далі радіус пошуку гіпотез недоцільно.

Перспективним напрямком подальшого вивчення питання побудови орфокоректорів на основі контекстно-асоціативних методів визначення варіантів виправлення є введення ранжування семантичних відношень лексико-семантичного словникового ресурсу за певними критеріями та врахування ваги відповідних дуг графа під час вибору його наступних вершин для аналізу.

Висновки

Доведено факт збереження точності, визначено умови покращення часових характеристик роботи коректора при перенесенні семантичного фільтра гіпотез на початок послідовності функцій підбору варіантів виправлення. Показано, що найкращі показники щодо швидкості роботи ПЗ забезпечує перенесення контекстно-асоціативної обробки текстових даних на етап висунення гіпотез.

Запропоновано інверсний контекстно-асоціативний метод виправлення орфографічних помилок шляхом ітеративного чергування процедур відбору гіпотез із словника за ознакою семантичної близькості до контексту та наступної їх перевірки на відповідність формальним критеріям подібності до спотвореного слова, такий, що забезпечує підвищення швидкості та точності роботи орфокоректора за рахунок зменшення потужності множини слів, які при цьому обробляються.

Запропоновано трактування точності корекції як відношення числа вибраних програмою вірних варіантів написання слова до загальної кількості отриманих варіантів, що дозволяє оцінювати точність виправлення помилок у випадку підбору гіпотез з великого за обсягом словника, а також фіксувати зміни у загальній кількості варіантів виправлення, запропонованих програмою.

Дослідження на практиці ефективності запропонованого методу орфокорекції підтвердило теоретичні положення, викладені у статті, та дозволило визначити рекомендовані значення параметрів алгоритму виправлення помилок, встановлення яких забезпечує ефективну роботу програмного орфокоректора.

Література

1. Kukich K. Techniques for Automatically Correcting Words in Text // ACM Computing Surveys. – 1992. – Vol. 24, № 4. – P. 377-439.
2. Лавошникова Э.К. О компьютерной коррекции «популярных» ошибок в текстах на русском языке // НТИ, сер. 2. – 2003. – № 9. – С. 28-34.
3. Михайлюк А.Ю., Заболотня Т.М. Комбінований метод виправлення орфографічних помилок у текстових даних // Вісник Хмельницького національного університету. – 2007. – № 2, Т. 2. – С. 21-26.
4. Kashyap R.L., Oommen B.J. Spelling correction using probabilistic methods // Pattern Recognit.Lett. – 1984. – Vol. 2, № 3. – P. 147-154.
5. Schaback J., Li F. Multi-Level Feature Extraction for Spelling Correction In Proc.of the IJCAI-2007 // Workshop on Analytics for Noisy Unstructured Text Data. – Hyderabad, India, 8 January 2007. – P. 79-86.
6. Пещак М.М. Нариси з комп'ютерної лінгвістики. – Ужгород: Закарпаття, 1999. – 199 с.
7. Trushkina Yu. Context-based Ranking of Suggestion for Spelling Correction // In Proc. of the RANLP 2005. – Borovetz, Bulgaria, 21 – 23 September 2005.
8. Andrew R. Golding A. Bayesian Hybrid Method for Context-Sensitive Spelling Correction // In Proceedings of the ACL Third Workshop on Very Large Corpora. – June 1995. – P. 39-53.
9. Марченко О.О. Алгоритми семантичного аналізу природномовних текстів: Дис. канд. фіз.-мат. наук: 01.05.01 / КНУ ім. Тараса Шевченка. – К., 2005.
10. Файн В.С., Рубанов Л.И. Машинное понимание текстов с ошибками. – М.: Наука, 1991. – 151 с.
11. Заболотня Т.М. Оптимізація процесу контекстноорієнтованої орфокорекції шляхом спрощення обчислення міри семантичної близькості слів // Проблеми інформатизації та управління: Зб. наук. праць. Випуск 3 (21). – К.: НАУ, 2007. – С. 55-59.

Т.Н. Заболотняя, А.Ю. Михайлюк, Е.С. Михайлюк

Инверсионный контекстно-ассоциативный метод автоматической орфокоррекции

Теоретически обусловлен и предложен инверсионный контекстно-ассоциативный метод автоматического исправления орфографических ошибок, который обеспечивает повышение скорости и точности работы соответствующего программного обеспечения. Дано определение показателя результативности функционирования орфокорректора – точности его работы. Показана эффективность использования предлагаемого метода для исправления орфографических ошибок в массиве гетерогенных словосочетаний по критериям скорости и точности коррекции.

Стаття надійшла до редакції 09.07.2008